All models of satellite-derived phenology are wrong, but some are useful: A case study from northern Australia

Nicolas Younes [a,b,*], Karen E. Joyce [a,b], Stefan W. Maier [a,c]

[a] *Centre for Tropical Environmental and Sustainability Science and School of Science and Engineering, James Cook University, QLD, Australia*
[b] *Centre for Tropical Water & Aquatic Ecosystem Research (TropWATER), James Cook University, Townsville, Australia*
[c] *Maitec, P.O. Box U19, Charles Darwin University, Darwin, NT 0815, Australia*

ARTICLE INFO

ABSTRACT

Satellite-derived phenology (or apparent phenology) is frequently used to illustrate changes in plant phenology (i.e. true phenology) and the effects of climate forcing. However, each study uses a different method to detect phenology. Plant phenology refers to the relationship between the life cycle of plants and weather and climate events. Phenology is often studied in the field, but recently studies have transitioned towards using satellite images to monitor phenology at the plot, country, and continental scales. The problem with this approach is that there is an ever-increasing variety of earth observation satellites collecting data with different spatial, spectral, and temporal characteristics. In this paper we ask if studies that detect phenology using different sensors over the same site produce comparable results. Mangrove forests are one example where different methods have been used to examine their apparent phenology. In general, plant phenology, including mangroves, is described using few individual plants, but continental-scale descriptions of phenological events are scarce or inexistent. Few attempts have been made to describe the phenology of mangroves using satellite imagery, and each study presents a different method. We hypothesize that apparent phenology changes with: 1) areal extent; 2) site location; 3) frequency of observation; 4) spatial resolution; 5) temporal coverage; and 6) the number of cloud contaminated observations. Intuitively, one would assume that these hypotheses hold true, yet few studies have investigated this. For example, one would expect that clouds change the observed phenology of vegetation, that the number of species captured at spatial resolution will impact the apparent phenology, or that mangroves in different places display different phenologies, but how are these changes represented in the apparent phenology? We use the Enhanced Vegetation Index (EVI) to examine the changes in the start of season and peak growing season dates, as well as the shape and amplitude of the apparent phenology in each hypothesis. We use Landsat and Sentinel 2 imagery over the mangrove forests in Darwin Harbour (Northern Territory, Australia) as a case study, and found that apparent phenology does change with the sensor, site, and cloud contamination. Importantly, the apparent phenology is comparable between Landsat and Sentinel 2 sensors, but it is not comparable to phenology derived from MODIS. This is due to differences in the spatial resolution of the sensors. Cloud contamination also significantly changes the apparent phenology of vegetation. In this paper we expose the complexity of modelling phenology with remote sensing and help guide future phenology investigations.

1. Introduction

Phenology is the branch of science that relates the life-cycle events of organisms to the biotic and abiotic events that cause them (Cerdeira et al., 2018); for example, plants losing their leaves in autumn and growing them in spring. Plant phenology is usually studied in the field. Field observations allow researchers to closely examine the life cycles of plants; however, these studies are often limited to small areas and to short periods of observation. In addition, historical records of plant phenology are limited in their geographical extent, number of species, and availability. Recently, studies have used satellite images to examine plant phenology at larger scales and over longer periods of time (Broich et al., 2014; Verbesselt et al., 2010; White et al., 2009). Here we use the term 'apparent phenology' to refer to satellite-derived phenology to discriminate it from the phenology observed in the field. The advantages of using satellites for studying phenology are threefold: 1) there are

---

decades worth of images for most places on earth; 2) satellites collect data at periodic intervals over large areas, and 3) all images are collected in the same way, thereby ensuring the data collection remains consistent over space and time. These advantages imply that we can assess phenology over large areas and retrospectively, however satellite images are just a tool to detect plant phenology, and, like any other tool, these images can be used in several ways.

From entire continents, to individual countries and small regions, satellite images are *the* preferred tool to detect plant phenology at different spatial scales (Czernecki et al., 2018; Pastor-Guzman et al., 2018; White et al., 2002). Like other tools, satellite images have different characteristics depending on the sensor that collects the data. Two important characteristics of satellite images that relate to the sensor are the spatial resolution and the temporal resolution. The spatial resolution refers to the land surface area covered by each pixel (i.e. pixel size), and the temporal resolution refers to the frequency with which the satellite visits a given site. Each sensor has a unique combination of spatial and temporal resolution, and the most frequently used sensors to detect plant phenology are Landsat, Sentinel 2, and MODIS (Moderate Resolution Imaging Spectroradiometer) (Bolton et al., 2020; Fu et al., 2014; Melaas et al., 2018). But detecting phenology not only depends on the sensor, it also depends on the amount of data available for the study site, and environmental factors such as cloud cover.

Many ecosystems are covered by clouds throughout the year, therefore obtaining cloud-free satellite images from those regions is rare. In other occasions, the target feature or phenomenon occurs when clouds are prevalent and, therefore, they are hidden from satellite view. When presented with these circumstances, researchers interested in phenology have resorted to using methods such as compositing to remove the effects of clouds in their data (e.g. (Griffiths et al., 2019)). Such approaches, however, may remove important within-year phenological features, making them less than ideal in many cases. More often than not, studies resort to masking algorithms to remove pixels marked as clouds, or cloud shadows; however, these algorithms are not perfect and can remove valid observations (i.e. commission errors), or keep invalid data (i.e. omission errors) (Ernst et al., 2018). The effects of clouds and cloud shadows on the satellite-derived (i.e. apparent) phenology of ecosystems are yet to be explored. One advantage of using cloud masking algorithms is that they remove unwanted data from the dataset, however this also reduced the number of usable observations to detect phenology.

The number of observations available for any given site is known as the time series, and it varies from sensor to sensor. Because detecting plant phenology requires (at least) one year of observations, the length of the time series plays an important role not only in detecting phenology, but also in evaluating changes over time (Garonna et al., 2016; Melaas et al., 2013; Vogelmann et al., 2016). Changes in plant phenology serve as indicators of alterations in weather and climate patterns (Garonna et al., 2016; Peñuelas et al., 2009; Sitch et al., 2015). However, there is no consensus on the most effective methods to detect these changes.

When referring to the ability of a model to represent the exact behaviour of a system, George Box famously stated the '*all models are wrong, but some are useful*'' (Box, 1979, 1976). This assertion seems to hold true for phenology investigation, where studies have used wide range of models across different ecosystems, spanning different decades, and using different sensors. Because of this variation in methods and inputs, comparing the phenology from two or more studies remains a challenging endeavour.

Vegetation phenology has been studied at different spatial scales and over different landscapes (Melaas et al., 2018; Pastor-Guzman et al., 2018). Broich et al. (2014) studied the phenology across Australia, and found stark differences between landscapes. Plant communities in central Australia, where rain is scarce, displayed small variations in greenness and, in some cases, seemed to skip the growing season altogether due to the lack of available water. The apparent (i.e. satellite-

derived) phenology showed by the authors corresponds, to some degree, to the real phenology of the vegetation. However, the variability in the vegetation within a 500 m pixel could be influencing the signal captured by the model. These, and many other models are wrong. They are wrong not because they lack merit or scientific validity, but because are incapable of representing the exact behaviour of the ecosystem.

But these models are also very useful. These models provide a good starting point for plant phenology investigations, but with an ever-increasing variety of earth observation satellites, it is important to raise the question: when two studies use different sensors to detect phenology over the same site, are their results comparable to one another? Our aim is to answer this question. To determine if satellite-derived phenology is comparable between studies, we evaluate if: 1) the size of the study site, 2) the location of the study site, 3) the frequency of observation, 4) the spatial resolution, 5) the temporal coverage, and 6) the data pre-processing methods change the apparent phenology of the study area. We test these hypotheses using mangrove forests as an example, but these considerations also apply to other intertidal ecosystems, inland wetlands, and tropical evergreen forests.

2. Methods

2.1. Study area

For this study we selected the mangroves that surround the city of Darwin, in Australia's Northern Territory because their phenology and distribution are well-known (Coupland et al., 2005; Metcalfe et al., 2011; Rogers et al., 2017). Darwin Harbour is comprised by three main sections or 'arms': the western arm, the central arm, and the eastern arm, which borders Darwin city. Within Darwin Harbour we selected one large area of the central arm (4500 ha, Fig. 1A), one smaller area covering 140 ha (Fig. 1B), and three small plots (1 ha each, Fig. 1B), as study sites. To avoid confusion, we will refer to them as 'region', 'zone', and 'plots' respectively.

Darwin harbour has an approximate area of 450 km2, out of which 190 km2 are covered by mangrove forests (Metcalfe et al., 2011). This region has a monsoonal climate and experiences maximum daily temperatures surrounding 32 °C all year long, while minimum temperatures vary between 18 °C and 25 °C. Precipitation is seasonal, with January, February and March being the months with highest rainfall, averaging 1105 mm; the drier months see around 8 mm. This region is also prone to tropical cyclones. According to the Australian Bureau of Meteorology Tropical Cyclone Data Portal (accessed 22/11/2020), since the 2009–2010 season over 14 tropical cyclones have been within a 300 km radius of Darwin Harbor, and several made landfall. Mangrove zonation in this region is highly dependent on tidal elevation due to an abrupt elevation 2 m rise from the coastal fringes. Mangrove diversity in Darwin harbour is high, with 48 mangrove species identified (Moritz-Zimmermann et al., 2002, p. 6); the species that dominate the seaward margin are mainly *Rhizophora stylosa* and *Sonneratia alba*, and *Ceriops australis* dominate the mid and upper tidal flats (Metcalfe, 1999, p. 16).

2.2. Data acquisition and processing

In this project we used 352 Landsat and 260 Sentinel 2 images, collected over our study area between January 2010 and January 2020. Landsat satellites collect data every 8–16 days in the visible and near infra-red regions of the electromagnetic spectrum at 30 m, while the Sentinel 2 sensors collect data every 5–10 days at 10 m resolution in similar spectral regions. Also, Landsat imagery for the Australian mainland goes back to 1987 but only to 2015 for Sentinel 2 sensors. All images were georegistered, atmospherically corrected, and grouped into tiles with a defined equal area projection by Geoscience Australia and loaded into the Digital Earth Australia platform (Dhu et al., 2017; Lewis et al., 2017; Roberts et al., 2017).Cloudy pixels, and missing pixels from Landsat 7 images (i.e. due to the Scan Line Corrector failure) were
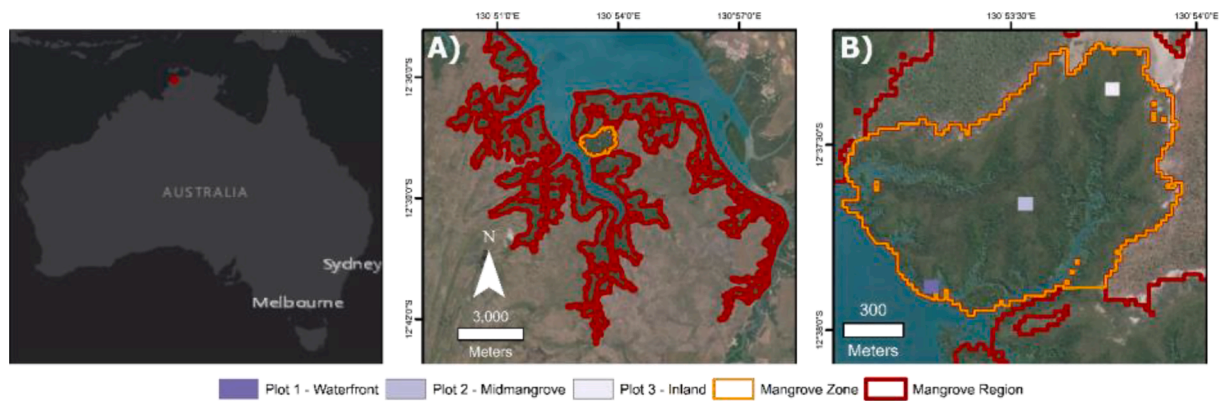
Fig. 1. Study area and selected sites. Panel A) shows the central arm of Darwin Harbour and the mangrove Region and mangrove Zone. Panel B) shows the mangrove Zone and the three mangrove plots (i.e. plot 1, plot 2, and plot 3). Basemap source: Esri. Mangrove polygon source: "Mangrove canopy cover" version 2.0.2 (Lymburner et al., 2019).
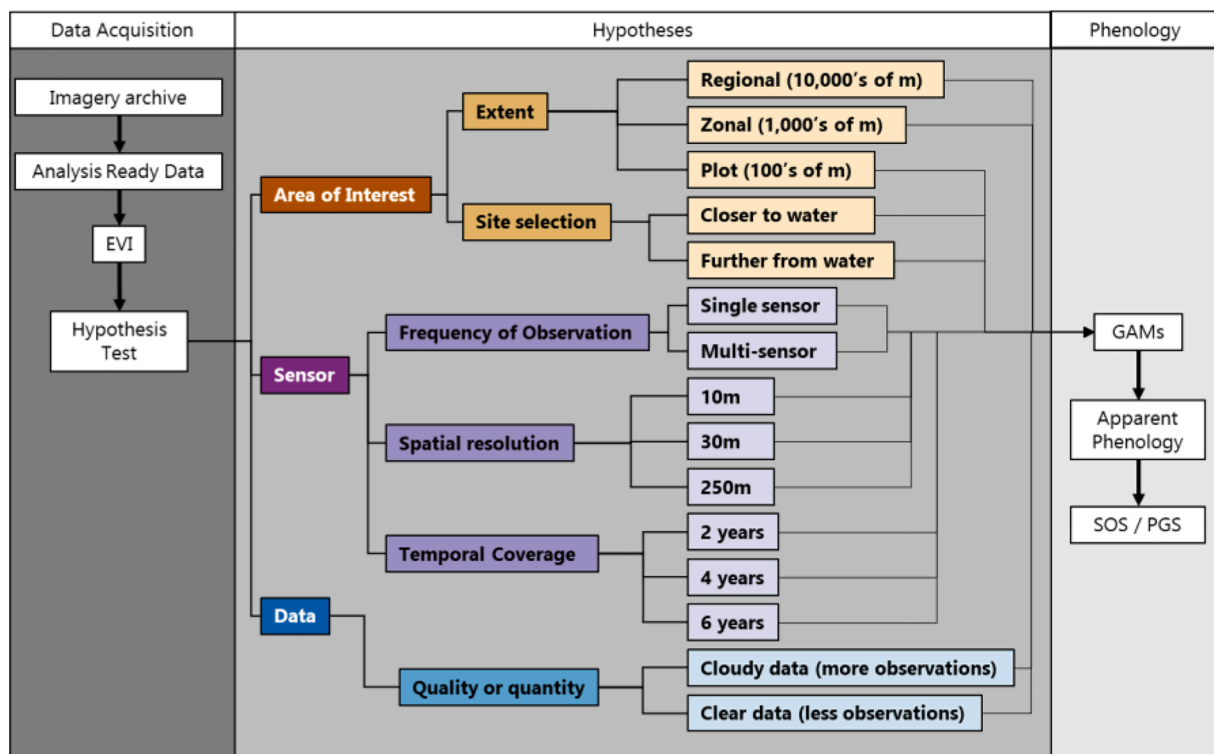
Fig. 2. Conceptual diagram of the hypotheses tested. From the apparent phenology we derived the following metrics: Start of Season (SOS) and Peak of Growing Season (PGS) represent the lowest and highest EVI values in the apparent phenology curve. Sentinel 2 data were resampled to obtain 250 m resolution imagery.

removed from the datasets and not used (i.e., no gap-filling was done). We used Digital Earth Australia to load all satellite images, aggregate the data, and remove all pixels that did not correspond to mangrove forests using the mangrove extent data from the National Map (https://nationalmap.gov.au/, dataset "Mangrove canopy cover" version 2.0.2). We calculated the Enhanced Vegetation Index (EVI) (Huete et al., 2002) for every mangrove pixel in every image; we selected EVI because it is known to be effective in mangrove and phenology investigations (Pastor-Guzman et al., 2018; Younes et al., 2019). To test hypothesis 4, we used Sentinel 2 images and resampled them to 30 m and 250 m to simulate the spatial resolution of the Landsat and MODIS sensors.

We used Generalized Additive Models (GAMs) to represent the phenology of our mangrove sites and to calculate two commonly-used metrics: Start of Season (SOS) and Peak of Growing Season (PGS). Importantly, while GAMs can be used to model phenology on a pixel by

pixel basis, for simplicity, we aggregated the data in the following way: first, we grouped all images by date of image acquisition, and then we extracted the median EVI of all pixels in the area under analysis (The result of this aggregation is a singe EVI value per site per date. We selected the geometric median because it produces statistically representative results, and can be used for time series investigations (Roberts et al., 2017).

2.3. Generalized additive models

To derive the apparent phenology of each study site, we used GAMS. Briefly, the principle behind GAMs is that the relationship between predictor and response variables is captured by functions instead of coefficients (Jones and Almond, 1992; Jones and Wrigley, 1995; Wood, 2017). In equation (1), a linear, additive, relationship is captured by the

coefficients β_1 and β_2. In contrast, GAMs capture the same additive relationship by using the functions $f_1(\cdot)$ and $f_2(\cdot)$. From polynomial and quadratic to complex thin-plate and Duchon splines, these functions can take many forms (Wood, 2017; Zuur et al., 2014).

$$Y = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon \tag{1}$$

$$Y = \beta_0 + f_1(X_{i1}) + f_2(X_{i2}) + \varepsilon \tag{2}$$

One advantage of using GAMs over other modelling techniques is that GAMs do not require the measurements (in this case EVI) to be evenly spaced (Taylor et al., 2018). This results very convenient because clouds, cloud shadows, and other errors cause random gaps in the data. These errors are not evenly distributed over space or time (e.g. Landsat 7 SLC error). The apparent phenology of each site was extracted from the satellite imagery using the 'Prophet' package (version 0.3.post2) for python (Taylor et al., 2018). This package was developed by Facebook to examine user engagement with the social media platform, however we repurposed it to detect intra- and interannual variations in EVI (i.e. phenology).

2.4. Data analysis

Vegetation phenology can be studied at the continental, regional, or plot level. Some studies look at vegetation phenology at the pixel level while others perform temporal or spatial aggregations to generate landscape averages (Czernecki et al., 2018; White et al., 2009). In this study we use three spatial scales (i.e. region, zone, plot) and aggregate the data for each spatial scale as shown in We mainly used the 'Xarray' (Hoyer and Hamman, 2017) and 'Prophet' (version 0.6, Taylor and Letham, 2018) packages for Python to create the models of apparent phenology for each analysis, and all model parameters remained unchanged throughout the analyses. As a result, any differences in the apparent phenology are derived from changes in the input data, not from changes in the model parameters. From the models of phenology, we derived the SOS and PGS days, defined as the median of the days of the year with the lowest and highest 5% of EVI values in the apparent phenology curve, respectively. In each hypothesis, we compared the following aspects of apparent phenology: number of peaks, number of troughs, median SOS date, median PGS date, maximum EVI, and minimum EVI, as well as the shape of the apparent phenology curve and the trend.

Importantly, we tested only one hypothesis at a time, meaning that there was no interaction between different areas, locations, frequencies of observation, spatial resolutions, time coverage, or cloud contamination. For most hypotheses we used all available Landsat images between 2010 and 2020, and cloud masked following Dhu et al. (2017) and Li et al. (2012). However, to test some hypotheses we used Sentinel 2 images, or a different time frame (see below). Furthermore, most scientific studies related to plant phenology rely on spectral indices such as EVI, rather than using individual spectral bands, therefore we do not investigate the effects of the spectral characteristics of the sensor on apparent phenology. We used Digital Earth Australia to test our six hypotheses, and below is a brief description of each analysis.

2.4.1. Hypothesis 1: Influence of size of area under investigation on apparent phenology.

To examine if the size of the study site changes the apparent phenology, we used all available Landsat images and compared models at the regional, zonal, and plot scales between the years 2010 and 2020 (Here we define regional, zonal, and plots, as spaces covering thousands of hectares, hundreds of hectares, and 10 ha or less respectively. To detect the apparent phenology, we used the median EVI of each site as input data for the GAMs.

2.4.2. Hypothesis 2: Influence of location on apparent phenology

When assessing the phenology of mangroves and other intertidal ecosystems, the site selection may play a crucial role in the results. To test our hypothesis, we compared three plots of the same size (3 × 2 pixels), located within 500 m of each other, as shown in One plot was located next to the high tide mark, one plot was located 300 m from the observable water edge, and the last plot was located (500 m from the observable water). Phenology was extracted from all Landsat 5/7/8 sensors for years 2010 to 2020.

2.4.3. Hypothesis 3: Influence of frequency of observation on apparent phenology

This hypothesis arises from the fact that each Landsat satellite (i.e. Landsat 7, and Landsat 8) has a temporal resolution of 16 days, but when combined, the temporal resolution of these sensors is reduced to eight days. Similarly, the Sentinel 2 sensors revisit frequency is 10 days, but only five when combined (Li and Roy, 2017). To investigate how the frequency of observations affects the observed phenology of mangrove forests, we compared the apparent phenology using all available observations from two individual sensors (i.e. Landsat 8 and Sentinel 2A), and of two families of sensors (i.e. Landsat and Sentinel 2) over the same period of time (i.e. 2015 – 2020). We selected these years because it is when both families of sensors have two operational satellites. All data was kept at native resolution, that is, 30 m for Landsat images, and 10 m for Sentinel 2 images. As a result, four models were generated: 1) Landsat 8; 2) Sentinel 2A; 3) Landsat 7/8; and 4) Sentinel 2A/B.

2.4.4. Hypothesis 4: Influence of spatial resolution on apparent phenology

We examined the effects of the spatial resolution on the apparent phenology of mangroves by using all available Sentinel 2 observations between 2010 and 2020. To simulate the spatial resolution of Landsat and MODIS sensors, we used analysis-ready Sentinel 2 images (Dwyer et al., 2018), we resampled the surface reflectance images to 30 m, and 250 m, and from the resampled images we derived the EVI. This resulted in three datasets for the selected area (i.e. 10 m, 30 m, and 250 m). Finally, we used GAMs to detect the apparent phenology.

2.4.5. Hypothesis 5: Influence of temporal coverage on apparent phenology

To examine how the length of a time series changes the apparent phenology of mangroves we used Landsat 8 observations of the study zone Fig. 1B) to create models spanning different lengths of time. To simulate different temporal coverages, we subset the Landsat 8 archive in the following way: one dataset spanning two years (2013–205), one dataset spanning four years (2013–2017), and one dataset using all data (2013–2020). We fitted the GAMs to each dataset and compared the results.

2.4.6. Hypothesis 6: Influence of cloud contamination on apparent phenology

To explore the extent to which apparent phenology is affected by cloud-contaminated observations, we compared two phenology models using all Landsat 5/7/8 observations over the mangrove region between 2010 and 2020 (Fig. 1A). As described by Lewis et al. (2017), Digital Earth Australia creates a 'Pixel Quality' band for each observation that results from testing each pixel for the presence of clouds or cloud shadows, radiometric saturation and zero values. This band allows the user to select or remove pixels that are relevant for their study. In our case, we used the 'Pixel Quality' band to remove observations tagged as clouds or cloud shadows from each image for the first model. For the second model, we did not remove the pixels tagged as clouds. Under the assumption that the 'Pixel Quality' band had few commission and omission errors, one model had fewer, but higher quality, mangrove observations, while the other model had more observations, but with lower quality overall.

3. Results

In Table 1 we show the median SOS and PGS, as well as the maximum

Table 1

Compilation of median SOS and PGS, maximum and minimum EVI for each hypothesis tested.

| Hypothesis	Site	Resolution (m)	Sensors	Median SOS	Median PGS	Maximum EVI	Minimum EVI
Hypothesis 1: Influence of size of area under investigation on apparent phenology	Region	30	Landsat 5/7/8	35	134	0.81	0.41
	Zone	30	Landsat 5/7/8	29	136	0.87	0.42
	Plot 2	30	Landsat 5/7/8	299	136	0.95	0.48
Hypothesis 2: Influence of location on apparent phenology	Plot 1	30	Landsat 5/7/8	35	123	0.97	0.39
	Plot 2	30	Landsat 5/7/8	299	136	0.95	0.48
	Plot 3	30	Landsat 5/7/8	299	136	0.90	0.41
Hypothesis 3: Influence of frequency of observations on apparent phenology	Plot 2	10	Sentinel 2A	306	110	0.96	0.48
	Plot 2	10	Sentinel 2A/B	302	115	0.93	0.51
	Plot 2	30	Landsat 8	288	131	0.94	0.45
	Plot 2	30	Landsat 7/8	35	134	0.81	0.41
Hypothesis 4: Influence of spatial resolution on apparent phenology	Zone	250	Sentinel 2A/B	297	112	0.86	0.44
	Zone	30	Sentinel 2A/B	26	116	0.84	0.38
	Zone	10	Sentinel 2A/B	27	116	0.83	0.38
Hypothesis 5: Influence of temporal coverage on apparent phenology	Plot 2 (2 years)	30	Landsat 8	56	136	1.16	0.39
	Plot 2 (4 years)	30	Landsat 8	299	139	1.04	0.45
	Plot 2 (6 years)	30	Landsat 8	296	134	1.01	0.45
Hypothesis 6: Influence of cloud contamination on apparent phenology	Region (with cloud mask)	30	Landsat 5/7/8	35	134	0.81	0.41
	Region (without cloud mask)	30	Landsat 5/7/8	33	144	0.76	0.20

and minimum EVI for each apparent phenology model in each hypothesis. We chose the midmangrove plot (i.e. Plot 2) to test several hypotheses because it is less influenced by water under the canopy, and edge effects from being close to other vegetation communities. These results suggest that mangroves around Darwin Harbour reach the PGS between Day of Year (DOY) 110 (April) and 144 (May), that is to say, within a 34-day window. The SOS is more variable. Some models show the SOS on DOY 288 (October) while others show it on DOY 56 (February). The shape of the apparent phenology also changes within, and between hypotheses, with some alterations being more evident than others. Fig. 3 shows the apparent phenology curves from all models created here, and highlights two that are more dissimilar from the rest. The trend (i.e. the general increase or decrease over time) of most models did not show significant changes, except in hypothesis 5 therefore, we rarely refer to it in the text. In the following sections we describe

the changes to the apparent phenology curve, the SOS and PGS arising from each hypothesis tested.

3.1. Extent and site selection

3.1.1. Hypothesis 1: Influence of site size on apparent phenology

The size of the site does change the apparent phenology, but, the models display some similarities. From Fig. 4 it is clear that the apparent phenology for the three sites has a single peak and two troughs. The apparent phenology of the sites shows that the PGS occurs on similar DOY. For the region, zone, and plot the PGS is at DOY 134, 136, and 136 respectively. According to the models, the median SOS for the region and the mangrove zone occurs early in the year (DOY 35 and 29 respectively), but the median SOS for the plot happens on DOY 299. This difference can also be seen in Fig. 4G. Another difference in the models is
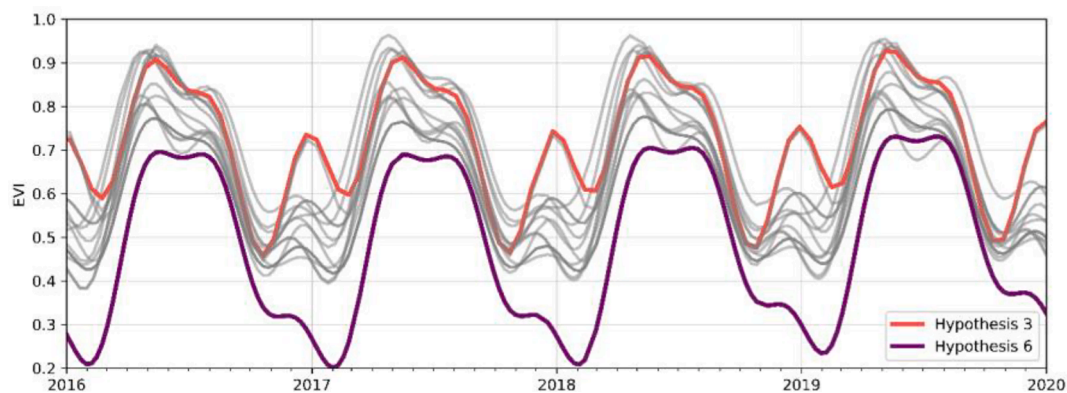
Fig. 3. Comparison of the apparent phenology models from all hypotheses (grey lines). Highlighted in red (Hypothesis 3) and purple (Hypothesis 6) are two models that show dissimilar patterns to other models. The trends of each model are not shown. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)
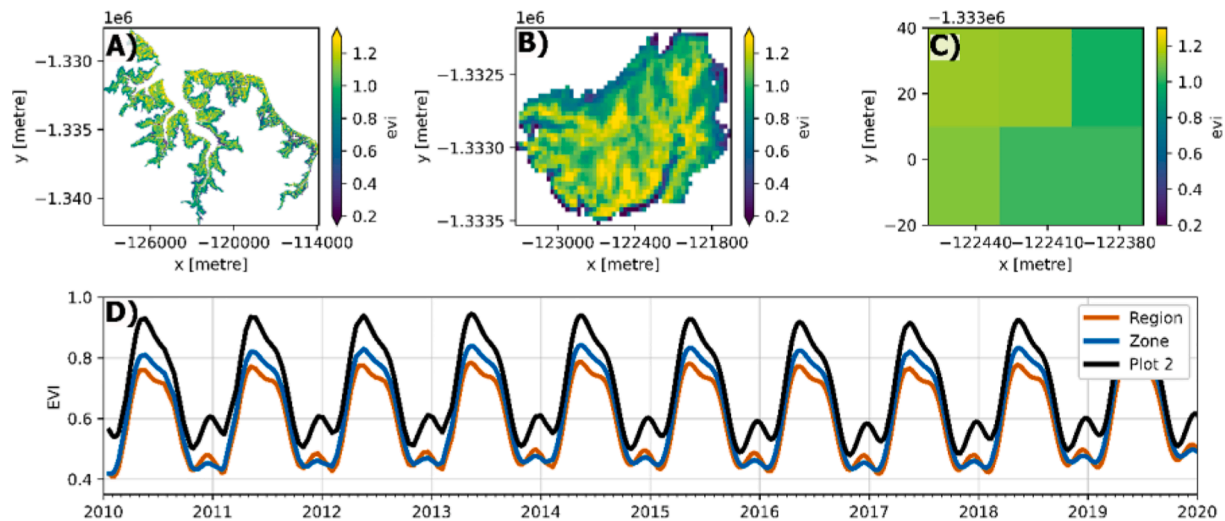
Fig. 4. Spatial representation and apparent phenology of mangroves at three different spatial scales: A) mangrove region, B) mangrove zone, C) mangrove plot 2 (midmangrove), and D) apparent phenology from the mangrove region, zone and plot 2. Note that the pixel size is the same across all samples. Median values are shown in all panels.

the range of EVI values. EVI ranges from 0.41 to 0.81 and 0.42–0.87 for the region and zone respectively, but those values change for the plot (i. e. 0.48–0.95).

3.1.2. Hypothesis 2: Influence of location on apparent phenology

Fig. 5 shows the apparent phenology for three plots located in three different intertidal zones, and, overall, the apparent phenology is similar across them. At the beginning of the calendar year, the three curves in Fig. 5 show a trough, followed by a well-defined peak; after this peak, the curves show a second trough followed by a smaller peak in EVI values.
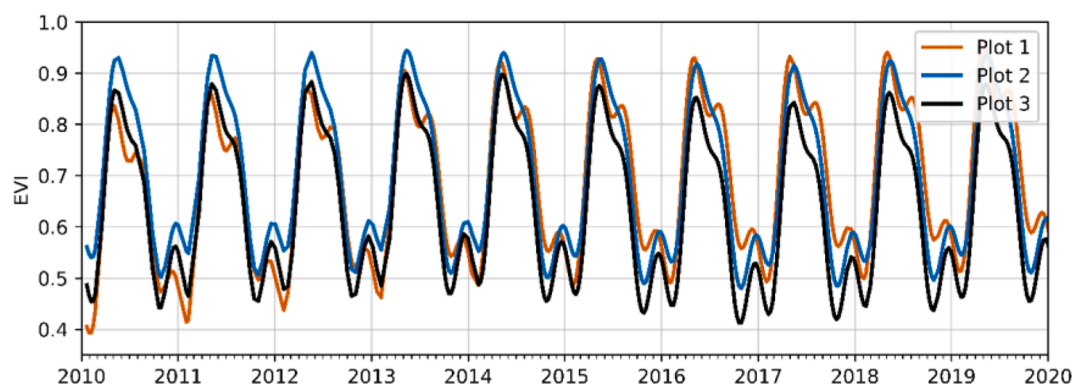
Mangroves in different intertidal zones display similar PGS dates, but different SOS dates. For example, the median PGS for plot 1 is 123, while the PGS for plot 2 and plot 3 is 136, and 136 respectively, resulting in a 13-day difference. In contrast, the median SOS occurs on DOY 35, 299, and 299 for plot 1, plot 2, and plot 3 respectively. For instance, the lines for plot 3 (black) and plot 2 (blue) in Fig. 5 display the lowest values in late 2010, while the line for plot 1 (orange) only shows a small trough. The lowest point of the phenology curve of plot 1 is reached in early 2011, as stated before, a recurring pattern between 2010 and 2020. Fig. 5 also shows that the EVI range for plot 1 changed from 0.39 to 0.86 in the 2010–2011 period, to 0.50–0.95 between 2018 and 2019. Plots 2 and 3 also display changes in their EVI range during different periods.

We do not attempt to analyze the factors leading to these variations in the EVI. Given that the model parameters to detect phenology of the three plots remained unchanged, our results support the idea PGS dates are comparable between plots, but the SOS dates are not comparable.

3.2. Sensor selection

3.2.1. Hypothesis 3: Influence of frequency of observation on apparent phenology

With a higher revisit time, there are approximately 51% more Sentinel 2A images of Darwin Harbour between 2015 and 2020 (n = 159), when compared to Landsat 8 (n = 105). Fig. 6 clearly shows that the frequency of observation changes the apparent phenology of plot 2, when comparing the apparent phenology from the Sentinel 2A and Landsat 8 sensors. The median PGS from Sentinel 2A occurs in DOY 110 (Fig. 6A), while the median PGS of the Landsat 8 model happens in DOY 131 (Fig. 6B). Regarding the SOS, the Sentinel 2A model shows that the lowest EVI values happen in DOY 306, while the phenology derived from only Landsat 8 images shows that the median SOS occurs in 288 (Fig. 6B). When comparing Fig. 6A to Fig. 6B, it is clear that there are differences in the shape of the curves. The Sentinel 2A model displays a single, well-defined peak in EVI values, while the Landsat 8 model shows two peaks, one in May, and one in December. This demonstrates that the

Fig. 5. Apparent phenology of three mangrove plots in different intertidal zones. Plot 1, Plot 2, and Plot 3 correspond to the waterfront, midmangrove, and inland plots respectively. The lines for plot 3 (black) and plot 2 (blue) sites have similar SOS and PGS dates over the years, but different EVI magnitudes. The apparent phenology of plot 1 (orange) has different SOS dates when compared to the other two sites. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)
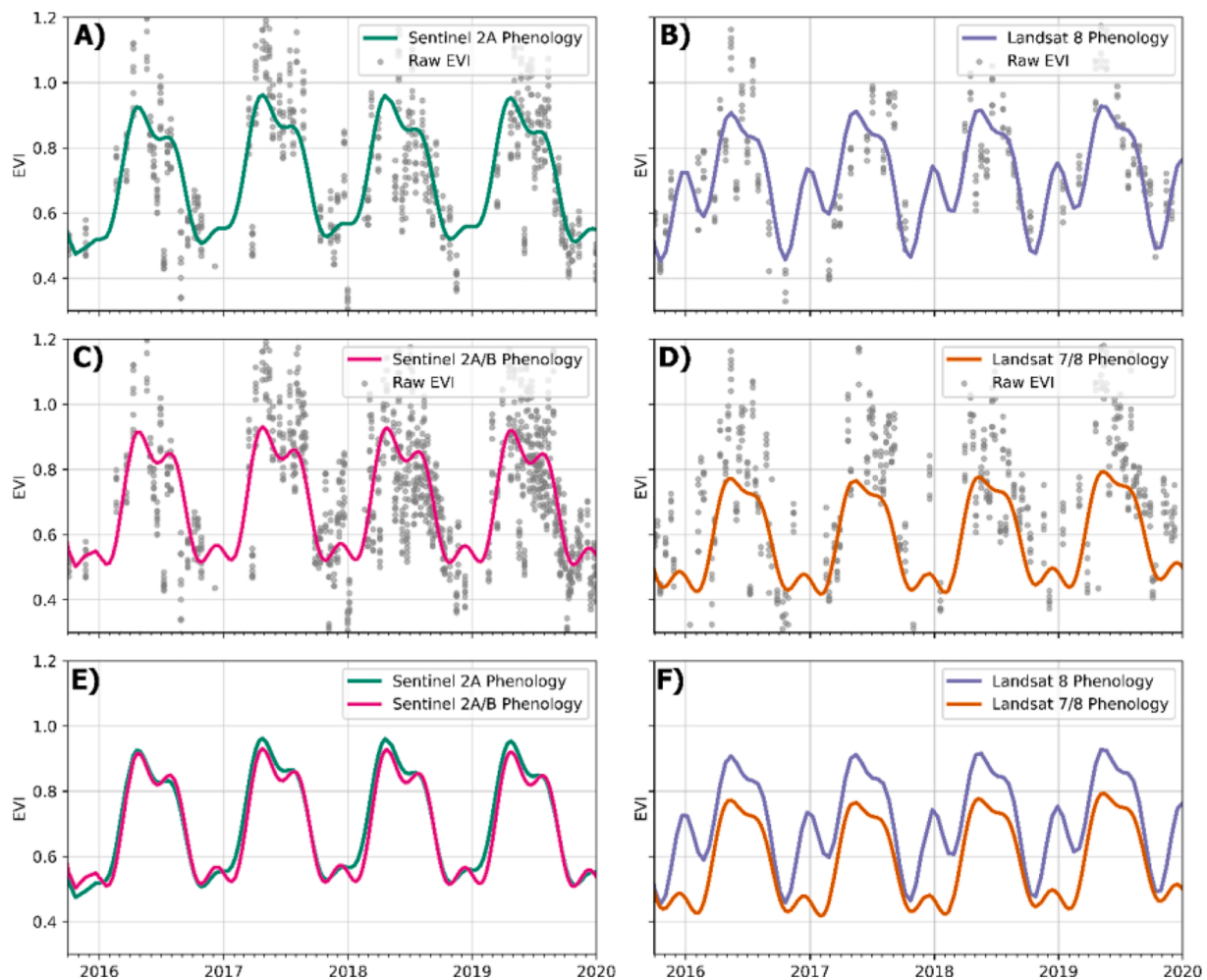
Fig. 6. Observation frequency and phenology from Sentinel 2A (A), Landsat 8 (B), Sentinel 2A/B (C), and Landsat 7/8 (D). Panels E) and F) compare the apparent phenology from the Sentinel and Landsat families for sensors respectively. For display purposes we only show the 2016–2020 period. Sentinel 2B was launched in September 2015.

frequency of observation does change the apparent phenology of mangroves.

When all Sentinel 2A/B images are considered (n = 260, Fig. 6C), the apparent phenology of plot 2 changes slightly when compared to the apparent phenology detected when only using Sentinel 2A images (Fig. 6C). The shape of both models is similar, while the EVI range varies slightly between the two models (0.48–0.96 for the Sentinel 2A, and 0.51–0.93 for the Sentinel 2A/B models). When the two Sentinel satellites gather information, the median PGS and SOS happen in DOY 115, and 302 respectively.

In sharp contrast to the Sentinel 2A/B case, when all Landsat 7/8 images are used (n = 207), the apparent phenology of plot 2 changes significantly when compared to the Landsat 8 phenology. Mainly, the December peak in EVI values changes to a trough, thereby moving the median SOS date from DOY 288 to DOY 35 (Fig. 6D/F). The median PGS experiences almost no change, occurring on DOY 134.

An important distinction between these models is the distribution of images over time. As shown in the Landsat 8 models (Fig. 6B), there are more observations between March and October, than between November and February each year. This results in a data void that the GAMs need to fill. The situation is similar for the Landsat 7/8 model (Fig. 6D) where, despite having double the observations, there are few data points between December and February. In contrast, the images used for the Sentinel 2A and Sentinel 2A/B model are more evenly distributed throughout the year, hence the apparent phenology resulting

from these models is more similar to those presented in the previous hypotheses.

3.2.2. Hypothesis 4: Influence of spatial resolution on apparent phenology

In Fig. 7 we demonstrate how the spatial resolution of the sensor changes the apparent phenology of the mangrove zone. The number of observations remains unchanged, therefore changes in the apparent phenology are related to the number of pixels and their EVI value in each scenario. For example, at the 10 m (Fig. 7A/B) and 30 m resolutions (Fig. 7C/D), the apparent phenology of the zone is almost identical. EVI ranges between 0.38 and 0.83, and 0.38–0.84 for the 30 m and 10 m models respectively, and the median SOS and PGS occur at approximately the same times: DOY 27 and 116 respectively for the 10 m model, and DOY 26 and 116 for the 30 m model (Fig. 7G). At these two resolutions, apparent phenology models are comparable with one another.

In contrast to the 10 m and 30 m resolutions, the EVI range for the 250 m resolution model ranges between 0.44 and 0.86, displaying a change in the amplitude of the model (Fig. 7E/F). The median PGS for the 250 m model occurs on DOY 112, however the median SOS occurs on DOY 297 (October), as opposed to January like in the finer resolution models. Taken together, these results suggest that the spatial resolution of the sensor can change the apparent phenology of an ecosystem. Therefore, images with similar resolutions (i.e. 10 m and 30 m) produce apparent phenologies that are comparable with one another.
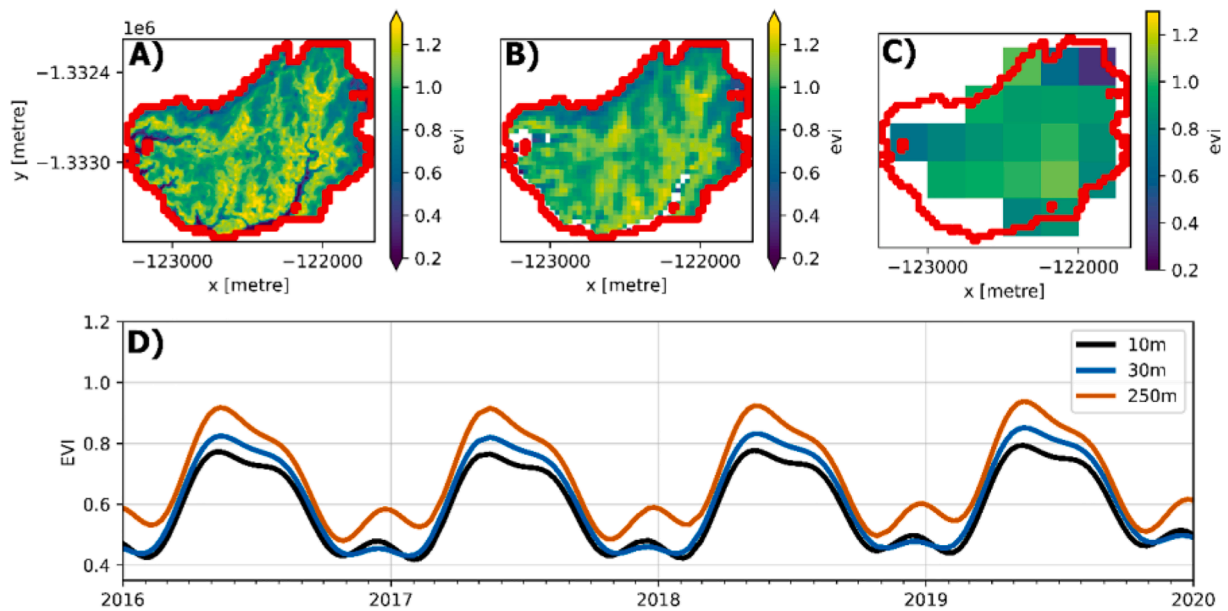
Fig. 7. Spatial representation and apparent phenology of the study zone from Sentinel 2 data. Panel A): Sentinel 2 data at native spatial resolution (10 m). Panel B) Sentinel 2 data resampled to match the spatial resolution of Landsat (30 m). Panel C): Sentinel 2 data resampled to match the spatial resolution of MODIS (250 m). Panel D): comparison of the apparent phenology of the study zone at three different spatial resolutions.

3.2.3. Hypothesis 5: Influence of temporal coverage on apparent phenology

The amount of data used by phenology studies alters the shape, amplitude, and apparent phenology metrics. Using Landsat 8 as an example, in Fig. 8 we demonstrate how the length of the time series changes the shape and amplitude of the apparent phenology, as well as changes in the trends in EVI values. Fig. 8A shows the apparent phenology of the two-year, four-year, and six-year models, as well as the raw EVI values (gray dots). It is easy to see that the range of EVI values for the two-year model of phenology is larger (0.39–1.16) than the EVI range for the four- and six-year models (0.45–1.04, and 0.45–1.01 respectively). Using only two years of observations (n = 38), the GAM tries to accommodate for all the peaks and troughs in the data, hence creating a model with a large amplitude. In this model, the median SOS and PGS occur on DOY 56 and 136 respectively.

With four years of data (n = 83), the EVI range is reduced, but the overall shape of the phenology is kept. The median PGS dates remain unchanged, but the median SOS moves to DOY 299. The model that used

six years of data (n = 136) is very similar to the four-year model. Maximum and minimum EVI values, including the intermediate peaks and troughs, do not show significant changes when compared to the four-year model, and the median SOS and PGS occur on DOY 296 and 134 respectively.

Regarding the trends in EVI values, the two- and four-year models show a downward trend, indicating that EVI values decrease each year, however, with only few years of data this should be interpreted with caution. The six-year model shows a downward trend EVI values between 2013 and early 2016, which is consistent with the previous two models. After 2017, the downward trend in the six-year model is replaced by an increasing trend, demonstrating that the temporal coverage alters the slope of the trend and the apparent phenology of mangrove forests.
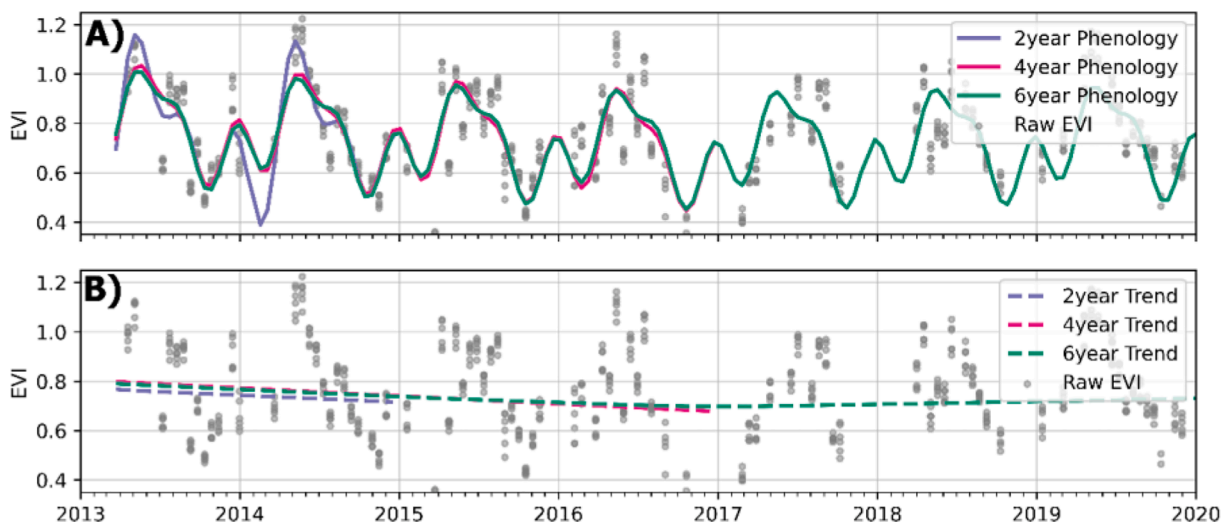
Fig. 8. Panel A) shows the apparent phenology curves using 2, 4, and 6 years of Landsat 8 observations. Panel B) shows the trends of apparent phenology from using 2, 4, and 6 years of Landsat 8 data.

3.3. Data and data processing

3.3.1. Hypothesis 6: Influence of cloud contamination on apparent phenology

We created two models of phenology using all Landsat 5/7/8 observations over the mangrove region between 2010 and 2020. The first difference between the models is the amount of data fed to the GAMs. As shown in Fig. 9A, the total number of usable pixels for the model *with* cloud masking is lower when compared to the number of usable pixels used for the model *without* cloud masking. The majority of the pixels masked by the algorithm have EVI values lower than 0.4, which correspond mostly, but not always, to clouds and cloud shadows. By removing these pixels the GAM has fewer input data, meaning that 1) there is less bias towards pixels with low EVI values, and 2) less variance in the data (see Section 3.3). The effects of reducing the bias and variances are clearly reflected in the apparent phenology (Fig. 9B).

With respect to the median SOS, the models display similar dates: DOY 34 and 33 for the models *with* and *without* cloud masking respectively (Fig. 9B). The median PGS however, shows a wider gap between the models. The model *with* cloud masking shows this event on DOY 134, while in the model *without* cloud masking the PGS occurs on DOY 144. From Fig. 9B it is easy to see that the amplitude of the model *without* cloud masking is greater (EVI between 0.20 and 0.76) than the amplitude of the apparent phenology model *with* cloud masking (EVI between 0.41 and 0.81).

4. Discussion

Studies that analyze plant phenology using satellite images are more common than ever before. However, these studies use a wide range of sensors, look at different sites, and process their data in different ways. The idea that "*all models are wrong, but some are useful*" (Box, 1979) is put to the test by asking: if two studies use different sensors to detect phenology over the same site, do they produce comparable results? We answer this question by testing six hypotheses related to the area of interest, the sensor selection, and the data pre-processing over an Australian mangrove forest (Fig. 2). Under certain circumstances, we found that studies may produce comparable phenology models, as explained below.

4.1. Area of interest (Extent and site Selection)

When studies perform spatial aggregation on sites that differ significantly in their areal extent (i.e. one order of magnitude or more), comparison is not advisable, mainly because of differences in the species composition, species richness, plant growth stages. All these factors can influence the shape and amplitude of the apparent phenology, as well as the SOS and PGS dates. Differences in SOS and PGS dates, as well as in EVI values support the idea that the size of the area of interest changes the apparent phenology of an ecosystem. Therefore, studies comparing sites of different areal extents may not be comparable with one another, even if data are spatially aggregated. Spatial aggregations are good for
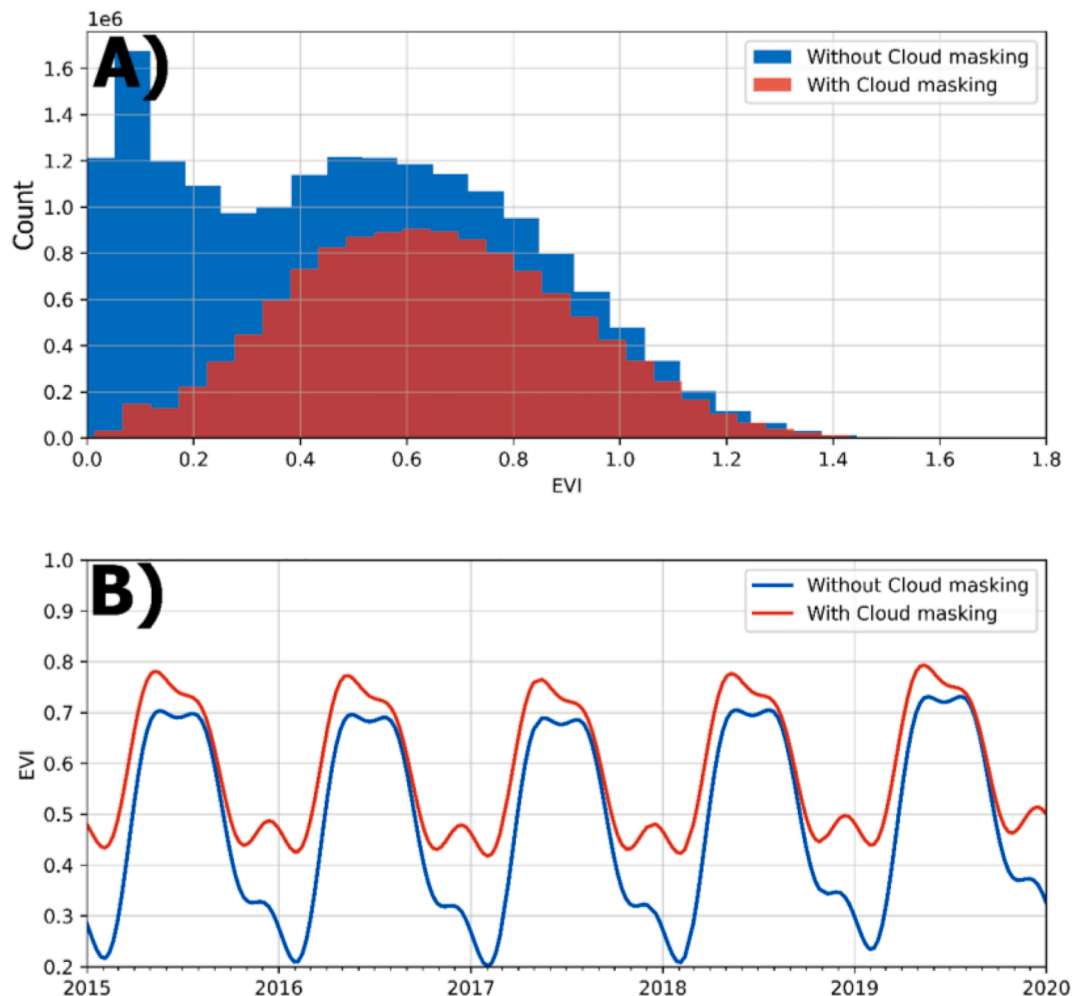
Fig. 9. Panel A) shows a histogram of the distribution of pixel values (i.e. EVI) used to detect the apparent phenology (2010–2020). Panel B) shows the apparent phenology of the mangrove zone when using data with cloud masking, and data without cloud masking.

making generalizations over selected areas, however they come with their own set of challenges (Jelinski and Wu, 1996; Wu and Li, 2009; Younes et al., 2019), including changing apparent phenology of vegetation. In our case, the apparent phenology curves were similar in shape, but the phenology metrics were different. The similarity in the shape of the three curves is not entirely surprising because 1) the region is dominated by mangroves of the *Ceriops* and *Rhizophora* genera (Metcalfe et al., 2011); 2) by using the median value of all pixels we capture the general trends in pixel values (Roberts et al., 2017); and 3) all models were created using the same number of input EVI values (i.e. the median EVI per date), which reduces the model variance. Despite these similarities, there are important differences in the apparent phenology of the mangrove region, zone, and plot.

Here, we have demonstrated that spatial aggregation of data results in changes to the SOS by more than 60 days and PGS. Our results agree with those of van Bussel et al. (2011), who conclude that spatial aggregations 1) can change the phenological metrics of crops by more than 10 days, and 2) they average the values of apparent phenology, thereby reducing natural variation and presenting a smoother phenology signal. Importantly, phenological metrics are key variables in assessing how an ecosystem responds to variations in the climate and, if these metrics are subject to change due to spatial averaging, they become unreliable, and potentially misleading.

Regarding the different locations, the models of apparent phenology shown here seem to coincide with the PGS dates, but not always with the SOS dates. These differences may be due to:

1. The species represented at the sites are slightly different and thus their phenology differs from one another. In other words, the dominant species in each site may change from site to site, therefore the phenology captured by the model differs.
2. The tidal height may play a role changing the apparent phenology of plot 1. Changes in apparent phenology may be due to variable tide heights, when the satellite image captures more water per pixel in plot 1 when compared to plots 2 and 3. Younes et al. (2019) described how different tidal heights might change the spectral signature of mangrove forests. The study concluded that, when using high spatial resolution imagery (1.4 m), the spectral signatures of mangrove forests does not change during high tide when compared to low tide. However, that study failed to examine the influence of the tides when a mixed pixel is more likely to occur. Further, Bishop-Taylor et al. (2019) demonstrated that optical sun-synchronous satellites fail to capture the entire tidal range in the Australian mainland. As a result, if the satellite consistently captures the higher tides, models of phenology at the fringes of mangroves may be biased towards lower values due to a larger proportion of water in each pixel (thus a smaller proportion of vegetation). Conversely, if the satellite captures the lower end of the tidal range, each pixel would have more vegetation exposed and thus, a more accurate phenology model.

Here, it is unclear if the tides have any real effect on the phenology model of plots 1 and 2, or if the mangrove species in each plot have slightly different phenologies. However, we demonstrated that while the phenology curves are similar between plots, the location of the site does play a role in the apparent phenology. If the areal extent of the sites remains similar, apparent phenologies should be comparable between studies, assuming other factors remain equal (i.e. sensor and data processing).

4.2. Sensor selection (frequency of observation, spatial resolution, and temporal coverage)

Understanding the needs of the phenological study is crucial to selecting the frequency of observations. For example, a study looking at crops may require a higher frequency of observations to examine the dates where leaf-growing occurs or when dormancy starts. In contrast, examining broad trends of leaf production and leaf fall in evergreen forests may require a lower frequency of observation. The frequency of observation is a recurring topic in phenology investigations. For example, Melaas et al. (2018) and Kowalski et al. (2020) take advantage of the frequency of observation of Landsat or Sentinel 2 sensors, but also the overlapping footprints of the images. It is easy to understand that, if there are more images available for a given time of year, the apparent phenology will be better represented than if there are fewer images. But if we have many images for one season (e.g. winter) and few, or none, for another (e.g. summer), the apparent phenology might not be well represented. In Section 2.2.1 we showed that, with more observations between September and April, the Landsat 7/8 model was more similar to the Sentinel 2 models than to the Landsat 8 model. With a higher temporal resolution, Sentinel 2 sensors have more useful (i.e. with no clouds) observations when compared to Landsat sensors, and this is directly related to the distribution of observations through time (Fig. 6). Here, we demonstrated that the frequency of observation changes the apparent phenology of a mangrove forest, but we also showed that the distribution of images over time is also an important factor. Recently, Kowalski et al. (2020) found that the frequency and distribution of observations changed the SOS dates throughout Germany, and Zhou et al. (2019) had similar findings in north American grasslands. Although their temporal and spatial scales differ from ours, our results align.

Turning to the spatial resolution of the sensor, to date, most studies on phenology have used sensors with a spatial resolution of 250 m, or coarser, particularly MODIS (see e.g. (Czernecki et al., 2018; White et al., 2009). This is too coarse to examine the phenology of a particular community, plot or species. It has been demonstrated the heterogeneous surfaces have different spectral responses at different spatial scales (Wu and Li, 2009), and that there are ways of minimizing this effect: either quantify the intra-pixel heterogeneity, or to use higher spatial resolution imagery (Younes et al., 2019). Considering that many studies of phenology use sensors with pixels that measure 250 m or more, it is inevitable to think that these models cannot accurately describe the phenology of a particular vegetation type. Using Box's words coupled with the findings by Wu and Li (2009), models that use large pixel sizes are more wrong than models that use smaller pixels sizes. Regardless of the pixel size, these models provide useful hints regarding the apparent phenology of the landscape.

Lately, the trend is to use Landsat or Sentinel 2 imagery individually or in combination. With higher spatial resolution than MODIS products, one would assume lower intra-pixel heterogeneity in Landsat and Sentinel 2 imagery, and thus, an apparent phenology that is less wrong. Despite this, a higher resolution means that one is not necessarily looking at the phenology of the landscape, but at the phenology of a plot or stand, turning this into a classic example of the modifiable area unit problem (Jelinski and Wu, 1996). Our results suggest that the spatial resolution of the sensor can change the apparent phenology of the ecosystem being investigated. Therefore, images with similar resolutions (i.e. 10 m and 30 m) produce apparent phenologies that are comparable with one another. Under the conditions tested here, the apparent phenology of our study zone is almost identical between the 10 m and 30 m resolution images. In other words, the models of apparent phenology from Landsat and Sentinel 2 sensors are comparable with one another.

In 2018, Pastick et al. combined Landsat and Sentinel 2 observations to detect the apparent phenology of dryland ecosystems in the western United States. To validate their models of phenology they compared the apparent phenology detected at 30 m to MODIS NDVI (Normalized Difference Vegetation Index) data collected at 250 m. At their native resolutions these two are not comparable, therefore the authors downscaled the 250 m data to 30 m, and they upscaled the 30 m data to 250 m. Similarly, when Claverie et al. (2018) describe the Harmonized Landsat and Sentinel dataset, they hint caution when comparing the 30 m data to the 250 m data. We too advise against the comparison of

phenology models from 10 m or 30 m to those of 250 m or coarser because of changes in the EVI range, and SOS dates.

Lastly, the effects of temporal coverage on apparent phenology are related to changes in the amplitude of the model (i.e. shape) and the SOS. Some studies look at plant phenology using a single year of data (Kowalski et al., 2020), but others examine the apparent phenology over decades (Garonna et al., 2016; Melaas et al., 2018). A single year of satellite images may give an idea of what the ecosystem is doing and when; however, examining one or two years of apparent phenology may not provide the full picture. Studies have demonstrated the difficulties in making generalizations of phenological cycles due to inter- and intra-annual variability in apparent phenology. For example Broich et al. (2014) used satellite imagery between 2000 and 2013 to examine apparent phenology across Australia. The authors conclude that, in some ecosystems of the Australian interior, the PGS varies by over a month from one year to the next, driven by rainfall patterns and atmospheric circulation. In comparison, our study site has a fairly regular phenology, however, the apparent phenology of the mangroves from two years of data had noticeable differences when compared to the four- and six-year models. Considering this, we advise caution when using less than four years of satellite imagery to detect apparent phenology, and we encourage incorporating contextual information in the analysis (e.g. weather and climate data) before comparing studies that span different periods.

Four years of data provide a more informative version of the phenology of the landscape than just one or two years, but it certainly does not allow for a comprehensive understanding of phenological events. As show by Broich et al. (2014) year-to-year variations in weather and climate patterns can, and often do, change the phenology of different vegetation types. In Australia, El Nino Southern Oscillation, the Julian-Madden Oscillation, and the Indian Ocean Dipole, control the amount of water contained in the atmosphere, and thus, precipitation over the continent. Precipitation, in turn, is a driver of vegetation phenology, making phenology susceptible to the aforementioned phenomena. Because one cannot assume that climate patterns are stable through time, one should not assume a stable phenology either. Under the conditions tested here, the shape of the four-year and six-year phenology curves and the long term trends are similar, but not identical. Feeding more data to the GAMs does change the output, and these differences should be tested with longer time series.

4.3. Data and data processing

We detected apparent phenology from a dataset that includes cloudy observations, and a dataset that removes most of the pixels affected by clouds and cloud shadows. Here, we demonstrate that the apparent phenology is, indeed, affected by clouds and cloud shadows (see also Section 3.5). While it is unlikely that many studies will use satellite images without any type of cloud removal, there are two main reasons why knowing how clouds affect apparent phenology is important: knowledge and context (Comber and Wulder, 2019).

Knowledge refers to understanding the *true* phenology of the study site, regardless of the vegetation type (e.g. crops, wetlands, or forests). For example, *when* the SOS and PGS happen and *how* they are expressed on the ground. Context, on the other hand, refers to the ability of the user to make informed decisions based on the apparent phenology. When knowledge and context are integrated, they allow the user to assess whether the apparent phenology is within an expected range of values (e.g. $0.4 < \text{EVI} < 1.2$) or if it affected by clouds or cloud shadows (e.g. $\text{EVI} < 0.4$). After comparing the true and apparent phenologies (i.e. using context and knowledge), the user can: 1) further clean the data, 2) tune or change the model, or 3) change the workflow (e.g. sensor fusion), to ensure that the apparent phenology is representative of the true phenology.

Moreover, many tropical ecosystems have prevalent cloud cover, which limits the number of clear observations in a year. Adding to this, the Landsat archive does not provide the same number of images over all continents (Wulder et al., 2016). Examples of areas that have persistent cloud cover and low number of images include the upper guinea forest in west Africa, the tropical rainforests of Central and South America. Due to the lack of clear observations, phenology investigations in these areas may resort to using cloud masks (or algorithms) with larger omission errors. With larger omission errors the probability of having more usable pixels for analysis increases, at the cost of alterations in the apparent phenology. Therefore, it is important to understand how clouds and cloud shadows change the apparent phenology of vegetation.

4.4. The definition of start of season

In this study we showed that the area and sensor(s) selected will change, to some degree, the shape, amplitude, and metrics of the apparent phenology. Our results indicate that the SOS is the metric most subject to change, and there are two main reasons for that: 1) the method used for representing apparent phenology, and 2) the definition of SOS. Here, we used GAMs to represent apparent phenology from satellite images, but most studies of phenology use fully parametric methods for this purpose. Studies that use fully parametric methods to represent phenology assume, among other things, that phenology can be represented by a model with a single peak and a single trough. This assumption may not hold true for mangroves and other evergreen forests, given that studies have demonstrated that some mangrove species display multimodal vegetative phenologies (Duke, 1990; Duke et al., 1984; Tomlinson, 1986). By using GAMs, we do not assume that there is a single peak or trough in the phenology of mangroves, and we let the data define the number, frequency and magnitude of the peaks and troughs. To prevent overfitting, we constrained the number of seasonal components in the 'Prophet' package. This means that, despite fitting a curve with fewer peaks and troughs, most models still display two peaks and two troughs in EVI values. Some of the changes in the SOS dates (e.g. in Section 2.1.1) may be due to the use of the median values for each date. By aggregating the data, we tend to capture the trend of the region or zone, in detriment of the natural variation of smaller plots. Given that we make no assumption about the number or magnitude of the peaks and troughs in the apparent phenology, PGS and SOS dates are subject to change (see Section 2).

The second reason that makes the SOS change is the way in which we defined it. We defined the SOS as the median of the lowest five percent of EVI values in a growth cycle. Others define SOS as the midpoint between the minimum spectral index value and the maximum green-up rate (Restrepo-Coupe et al., 2015). Broich et al. (2014) defined the SOS as the date where the EVI values reached 20% of the phenology curve amplitude, and tested their method on several vegetation types across Australia. Before fitting a double logistic curve to the data, the authors used a Savitzky–Golay filter to reduce the variability in EVI values. In doing so, the authors alter the data to fit a unimodal model of phenology, and therefore, the SOS dates. Under the conditions tested here, neither of the previous definitions may be adequate given that the data shows two periods of low EVI with an intermediate peak of varying magnitude. A new definition of SOS may be required to accommodate for unimodal and multimodal phenologies in different environments.

4.5. Uncertainties and limitations

In this document, we used Darwin Harbor as a case study for examining the influence that the site, sensor, and data have on apparent phenology. While we have tried to isolate every element to demonstrate its influence, there are some limitations that need to be acknowledged.

Firstly, we may not be looking at the same site every time. Most methods for geocoding satellite images are subject to errors of up to half a pixel, therefore using Sentinel or Landsat images implies that each pixel might be off target by up to 6 and 18 m respectively (Claverie et al., 2018). If, like here, one is aggregating the data before applying the

model, these errors might be reduced because the aim is to capture the phenology of an area larger than one pixel. Conversely, if the aim is to examine the phenology of every pixel, geo-location errors must be acknowledged, and cautious interpretation is advised.

Secondly, no cloud-masking algorithm is perfect. There will be omission and commission errors associated with each algorithm, and interpretation of phenology models needs to accommodate for such errors. Despite new methods being developed for cloud and cloud shadow detection for Landsat and Sentinel datasets (Candra et al., 2020; Ernst et al., 2018; Mateo-García et al., 2018), the user must acknowledge that abnormally low or high values may result from omitted clouds or shadows. In our case, some omission and commission errors were evident in the Landsat images despite masking the clouds and cloud shadows. We reduced these errors by aggregating the data, and using the median values for each date. Nevertheless, the user may decide not to do any spatial aggregations, and must acknowledge the limitations of cloud-masking algorithms. But pixel values are not only affected by clouds or shadows, they can be affected by vegetation migration.

Mangroves are dynamic ecosystems; they are known to migrate in response to changes in environmental conditions (Asbridge et al., 2015; Rogers et al., 2017; Saintilan et al., 2014). This implies that models of phenology may not only reflect annual cycles in the trees, but also plant migration, colonization, or dieback. Recently, Lymburner et al. (2019) mapped the extent of mangrove forests across Australia on a yearly basis. This information, coupled with high-resolution imagery, pheno-cams, and other ancillary datasets may provide insights as to the processes driving the phenology of a particular pixel, plot, or region (Moore et al., 2016).

While the claims that satellites collect data consistently over space and time are true, modifications in the sensor characteristics and orbit may show as changes in phenology. For example, Roy et al. (2016), and Zhang and Roy (2016) attributed some changes in surface reflectance to variations in orbits, sensing geometries, and spectral characteristics of Landsat 5, Landsat 7 and Landsat 8 sensors. Considering that these changes are small, but potentially significant, data harmonization should be considered when analysing a time series from these sensors.

There are also limitations related to GAMs, for example, their tendency to overfit the data. To overcome this, 'Prophet' and other packages offer the ability to limited the number of knots and penalizing models in favour of simpler ones. As a result, our models captured the principal features of the phenology without overfitting the data. There are, however, ways to improve upon the method used here. For example, combining the advanced capabilities of the 'mgcv' package to select different smooth functions, and Bayesian stochastic simulation (Wood, 2017).

Lastly, we used GAMs as a way to examine phenology from satellite imagery, while others have used fully parametric methods (Pasquarella et al., 2018; Verbesselt et al., 2010). Because the mechanisms behind the models differ from one another, phenology models may not be directly comparable unless the site, sensor, and data are the same (e.g. Atkinson et al. (2012)). Here we demonstrated the usefulness of GAMs as a tool for detecting mangrove phenology in Northern Australia, however our intention is not to compare this method to other methods or sites. Indeed, a comparison between parametric and semi-parametric methods is needed, not only in the context of assessing phenology, but also to determine if different sites and sensors alter the apparent phenology equally. Atkinson et al. (2012) compared several fully parametric algorithms used to detect phenology, but did not incorporate GAMs or other semi-parametric models. Change detection algorithms are often based on the vegetation changes over time (i.e. phenology), however they are rarely used for phenology investigations. Yan et al. (2019), for example, compared 'Prophet' (the algorithm used here) to other, commonly-used, algorithms for land cover change detection. The authors found that the algorithm that used 'Prophet' was able to accurately capture subtle changes in land cover, as well as the direction and rate of change. The performance of GAMs could be compared to change

detection models (Yan et al., 2019) especially in the context of climate forcing and its effects on phenological cycles.

4.6. Future directions

To date, most remote sensing studies on phenology have used MODIS, with fewer studies using Landsat or Sentinel data. Currently, there are several methods to harmonize Landsat and Sentinel-2 data to study phenology, however they focus mainly in dryland ecosystems (Pastick et al., 2018; Zhou et al., 2019). Data harmonization of these two families of sensors will improve phenology investigations in several ways, including: 1) near real-time monitoring of phenology, 2) increased frequency of observations, 3) more evenly distributed usable observations throughout the year, and 4) increased number of cloud-free observations in mangrove and other evergreen forests. With this wealth of methods and data, phenology investigations can move from examining past phenology to forecasting future phenology. Forecasting phenology will allow us to assess future carbon budgets, and select the ecosystems that will help mitigate the effects of increased concentration of carbon dioxide in the atmosphere.

5. Conclusions

In this paper we ask if studies that detect phenology using different sensors on the same site produce comparable results, and often the result is: no. We explored different area sizes, using one versus many sensors, and removing clouds. Each maintained some characteristics of the apparent phenology, but not all. Our intention is not to prescribe a method for modelling phenology, but to expose the complexity of the process and help guide future phenology investigations. We conclude that: 1) the areal extent and the location of the study site change the apparent phenology of the vegetation; 2) the distribution of observations over time may play a more important role on apparent phenology than the frequency of observation; 3) the apparent phenology resulting from Landsat and Sentinel 2 sensors is comparable with one another, but it is not comparable to phenology derived from MODIS due to differences in the spatial resolution of the sensors; 4) the length of the time series has more impact on the trend of the phenology than on the apparent phenology itself; and 5) cloud contamination significantly changes the apparent phenology of vegetation, therefore, studies must incorporate knowledge and context into their workflows to ensure apparent phenology resembles true vegetation phenology. Reflecting on what George Box said, the models tested here are able to represent behaviour of a mangrove ecosystem and, like many other models of phenology, the ones presented here are wrong. They are wrong because we are not certain about the miniscule and specific changes that take place in the ecosystem, or when they take place, or if they take place at the same time throughout the landscape. But the models are certainly useful. The usefulness of these models comes from their ability to represent data and its variations. From these models we can examine seasonal variations and long term trends, and eventually compare those variations to other biophysical drivers. In conclusion, each combination of sensor, site, and data will produce a unique apparent phenology model, therefore, all models of apparent phenology are wrong, but some are useful.

CRediT authorship contribution statement

Nicolas Younes: Conceptualization, Methods, Data analysis, Writing, reviewing and editing. **Karen E. Joyce:** Conceptualization, Methods, Reviewing and editing. **Stefan W. Maier:** Conceptualization, Methods, Reviewing and editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial

interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

References

Asbridge, E., Lucas, R., Accad, A., Dowling, R., 2015. Mangrove response to environmental changes predicted under varying climates: Case studies from Australia. Curr. For. Reports 1, 178–194. https://doi.org/10.1007/s40725-015-0018-4.

Atkinson, P.M., Jeganathan, C., Dash, J., Atzberger, C., 2012. Inter-comparison of four models for smoothing satellite sensor time-series data to estimate vegetation phenology. Remote Sens. Environ. 123, 400–417. https://doi.org/10.1016/j.rse.2012.04.001.

Bishop-Taylor, R., Sagar, S., Lymburner, L., Beaman, R.J., 2019. Between the tides: Modelling the elevation of Australia's exposed intertidal zone at continental scale. Estuar. Coast. Shelf Sci. https://doi.org/ 10.1016/j.ecss.2019.03.006.

Bolton, D.K., Gray, J.M., Melaas, E.K., Moon, M., Eklundh, L., Friedl, M.A., 2020. Continental-scale land surface phenology from harmonized Landsat 8 and Sentinel-2 imagery. Remote Sens. Environ. 240, 111685 https://doi.org/10.1016/j.rse.2020.111685.

Box, G.E.P., 1979. Robustness in the strategy of scientific model building. In: Robustness in Statistics. Academic Press, pp. 201–236.

Box, G.E.P., 1976. Science and Statistics. J. Am. Stat. Assoc. 71, 791–799. https://doi.org/10.1080/01621459.1976.10480949.

Broich, M., Huete, A., Tulbure, M.G., Ma, X., Xin, Q., Paget, M., Restrepo-Coupe, N., Davies, K., Devadas, R., Held, A., 2014. Land surface phenological response to decadal climate variability across Australia using satellite remote sensing. Biogeosciences 11, 5181–5198. https://doi.org/10.5194/bg-11-5181-2014.

Candra, D.S., Phinn, S., Scarth, P., 2020. Cloud and cloud shadow masking for Sentinel-2 using multitemporal images in global area. Int. J. Remote Sens. 41, 2877–2904. https://doi.org/10.1080/01431161.2019.1697006.

Cerdeira, M.L.P., Katharine, A., Irene, M., 2018. Rethinking tropical phenology: insights from long-term monitoring and novel analytical methods. Biotropica 50, 371–373. https://doi.org/10.1111/btp.12562.

Claverie, M., Ju, J., Masek, J.G., Dungan, J.L., Vermote, E.F., Roger, J.-C., Skakun, S.V., Justice, C., 2018. The Harmonized Landsat and Sentinel-2 surface reflectance data set. Remote Sens. Environ. 219, 145–161. https://doi.org/10.1016/j.rse.2018.09.002.

Comber, A., Wulder, M., 2019. Considering spatiotemporal processes in big data analysis: Insights from remote sensing of land cover and land use. Trans. GIS. https://doi.org/10.1111/tgis.12559.

Coupland, G.T., Paling, E.I., McGuinness, K.A., 2005. Vegetative and reproductive phenologies of four mangrove species from northern Australia. Aust. J. Bot. 53, 109–117. https://doi.org/10.1071/BT04066.

Czernecki, B., Nowosad, J., Jabłońska, K., 2018. Machine learning modeling of plant phenology based on coupling satellite and gridded meteorological dataset. Int. J. Biometeorol. 62, 1297–1309. https://doi.org/10.1007/s00484-018-1534-2.

Dhu, T., Dunn, B., Lewis, B., Lymburner, L., Mueller, N., Telfer, E., Lewis, A., McIntyre, A., Minchin, S., Phillips, C., 2017. Digital earth Australia – unlocking new value from earth observation data. Big Earth Data 1, 64–74. https://doi.org/10.1080/20964471.2017.1402490.

Duke, N.C., 1990. Phenological Trends with Latitude in the Mangrove Tree Avicennia Marina. J. Ecol. 78, 113. https://doi.org/10.2307/2261040.

Duke, N.C., Bunt, J.S., Williams, W.T., 1984. Observations on the Floral and Vegetative Phenologies of North-Eastern Australian Mangroves. Aust. J. Bot. 32, 87–99.

Dwyer, J.L., Roy, D.P., Sauer, B., Jenkerson, C.B., Zhang, H.K., Lymburner, L., 2018. Analysis Ready Data: Enabling Analysis of the Landsat Archive. Remote Sens. 10.

Ernst, S., Lymburner, L., Sixsmith, J., 2018. Implications of Pixel Quality Flags on the Observation Density of a Continental Landsat Archive. Remote Sens. https://doi.org/10.3390/rs10101570.

Fu, Y., Zhang, H., Dong, W., Yuan, W., 2014. Comparison of Phenology Models for Predicting the Onset of Growing Season over the Northern Hemisphere 9. https://doi.org/10.1371/journal.pone.0109544.

Garonna, I., de Jong, R., Schaepman, M.E., 2016. Variability and evolution of global land surface phenology over the past three decades (1982–2012). Glob. Chang. Biol. 22, 1456–1468. https://doi.org/10.1111/gcb.13168.

Griffiths, P., Nendel, C., Hostert, P., 2019. Intra-annual reflectance composites from Sentinel-2 and Landsat for national-scale crop and land cover mapping. Remote Sens. Environ. 220, 135–151. https://doi.org/10.1016/j.rse.2018.10.031.

Hoyer, S., Hamman, J.J., 2017. xarray: N-D labeled Arrays and Datasets in Python. J. open Res. Softw. 5 https://doi.org/10.5334/jors.148.

Huete, A., Didan, K., Miura, T., Rodriguez, E., Gao, X., Ferreira, L., 2002. Overview of the radiometric and biophysical performance of the MODIS vegetation indices. Remote Sens. Environ. 83, 195–213. https://doi.org/10.1016/S0034-4257(02)00096-2.

Jelinski, D.E., Wu, J., 1996. The modifiable areal unit problem and implications for landscape ecology. Landsc. Ecol. 11, 129–140. https://doi.org/10.1007/BF02447512.

Jones, K., Almond, S., 1992. Moving out of the Linear Rut: The Possibilities of Generalized Additive Models. Trans. Inst. Br. Geogr. 17, 434. https://doi.org/10.2307/622709.

Jones, K., Wrigley, N., 1995. Generalized Additive Models, Graphical Diagnostics, and Logistic Regression. Geogr. Anal. 27, 1–18. https://doi.org/10.1111/j.1538-4632.1995.tb00333.x.

Kowalski, K., Senf, C., Hostert, P., Pflugmacher, D., 2020. Characterizing spring phenology of temperate broadleaf forests using Landsat and Sentinel-2 time series. Int. J. Appl. Earth Obs. Geoinf. 92, 102172 https://doi.org/10.1016/j.jag.2020.102172.

Lewis, A., Oliver, S., Lymburner, L., Evans, B., Wyborn, L., Mueller, N., Raevksi, G., Hooke, J., Woodcock, R., Sixsmith, J., Wu, W., Tan, P., Li, F., Killough, B., Minchin, S., Roberts, D., Ayers, D., Bala, B., Dwyer, J., Dekker, A., Dhu, T., Hicks, A., Ip, A., Purss, M., Richards, C., Sagar, S., Trenham, C., Wang, P., Wang, L.-W., 2017. The Australian Geoscience Data Cube — Foundations and lessons learned. Remote Sens. Environ. 202, 276–292. https://doi.org/10.1016/j.rse.2017.03.015.

Li, F., Jupp, D.L.B., Thankappan, M., Lymburner, L., Mueller, N., Lewis, A., Held, A., 2012. A physics-based atmospheric and BRDF correction for Landsat data over mountainous terrain. Remote Sens. Environ. 124, 756–770. https://doi.org/10.1016/j.rse.2012.06.018.

Li, J., Roy, D.P., 2017. A Global Analysis of Sentinel-2A, Sentinel-2B and Landsat-8 Data Revisit Intervals and Implications for Terrestrial Monitoring. Remote Sens. https://doi.org/10.3390/rs9090902.

Lymburner, L., Bunting, P., Lucas, R., Scarth, P., Alam, I., Phillips, C., Ticehurst, C., Held, A., 2019. Mapping the multi-decadal mangrove dynamics of the Australian coastline. Remote Sens. Environ. 238, 111185 https://doi.org/10.1016/j.rse.2019.05.004.

Mateo-García, G., Gómez-Chova, L., Amorós-López, J., Muñoz-Marí, J., Camps-Valls, G., 2018. Multitemporal Cloud Masking in the Google Earth Engine. Remote Sens. https://doi.org/10.3390/rs10071079.

Melaas, E.K., Friedl, M.A., Zhu, Z., 2013. Detecting interannual variation in deciduous broadleaf forest phenology using Landsat TM/ETM+ data. Remote Sens. Environ. 132, 176–185. https://doi.org/10.1016/j.rse.2013.01.011.

Melaas, E.K., Sulla-Menashe, D., Friedl, M.A., 2018. Multidecadal Changes and Interannual Variation in Springtime Phenology of North American Temperate and Boreal Deciduous Forests. Geophys. Res. Lett. 45, 2679–2687. https://doi.org/10.1002/2017GL076933.

Metcalfe, K., 1999. Mangrove Litter Production in Darwin Harbour, NT. A study of litter fall as a measure of primary productivity in the mangrove communities of Darwin Harbour. Masters Thesis. Northern Territory University. http://espace.cdu.edu.au/view/cdu:38880.

Metcalfe, K.N., Franklin, D.C., McGuinness, K.A., 2011. Mangrove litter fall: Extrapolation from traps to a large tropical macrotidal harbour. Estuar. Coast. Shelf Sci. 95, 245–252. https://doi.org/10.1016/j.ecss.2011.09.006.

Moore, C.E., Brown, T., Keenan, T.F., Duursma, R.A., Van Dijk, A.I.J.M., Beringer, J., Culvenor, D., Evans, B., Huete, A., Hutley, L.B., Maier, S., Restrepo-Coupe, N., Sonnentag, O., Specht, A., Taylor, J.R., Van Gorsel, E., Liddell, M.J., Lawrence Berkeley National Lab. (LBNL) CA (United States), B., 2016. Reviews and syntheses: Australian vegetation phenology: New insights from satellite remote sensing and digital repeat photography. Biogeosciences. https://doi.org/10.5194/bg-13-5085-2016.

Moritz-Zimmermann, A., Comely, B., Lewis, D., Environment., D. of I.P. and, 2002. Darwin harbour mangrove monitoring methodology. Northern Territory Government, Department of Infrastructure Planning and Environment.

Pasquarella, V.J., Holden, C.E., Woodcock, C.E., 2018. Improved mapping of forest type using spectral-temporal Landsat features. Remote Sens. Environ. 210, 193–207. https://doi.org/10.1016/j.rse.2018.02.064.

Pastick, J.N., Wylie, K.B., Wu, Z., 2018. Spatiotemporal Analysis of Landsat-8 and Sentinel-2 Data to Support Monitoring of Dryland Ecosystems. Remote Sens. https://doi.org/10.3390/rs10050791.

Pastor-Guzman, J., Dash, J., Atkinson, P.M., 2018. Remote sensing of mangrove forest phenology and its environmental drivers. Remote Sens. Environ. 205, 71–84. https://doi.org/10.1016/j.rse.2017.11.009.

Peñuelas, J., Rutishauser, T., Filella, I., 2009. Phenology Feedbacks on Climate Change. Science (80-.). 324, 887 LP – 888.

Restrepo-Coupe, N., Huete, A., Davies, K., 2015. Satellite Phenology Validation, in: Held, A., Phinn, S., Soto-Berelov, M., Jones, S. (Eds.), AusCover Good Practice Guidelines: A Technical Handbook Supporting Calibration and Validation Activities of Remotely Sensed Data Product. TERN AusCover, pp. 155–157.

Roberts, D., Mueller, N., McIntyre, A., 2017. High-Dimensional Pixel Composites from Earth Observation Time Series. IEEE Trans. Geosci. Remote Sens. 55, 6254–6264. https://doi.org/10.1109/TGRS.2017.2723896.

Rogers, K., Lymburner, L., Salum, R., Brooke, B.P., Woodroffe, C.D., 2017. Mapping of mangrove extent and zonation using high and low tide composites of Landsat data. Hydrobiologia 803, 49–68. https://doi.org/10.1007/s10750-017-3257-5.

Roy, D.P., Kovalskyy, V., Zhang, H.K., Vermote, E.F., Yan, L., Kumar, S.S., Egorov, A., 2016. Characterization of Landsat-7 to Landsat-8 reflective wavelength and normalized difference vegetation index continuity. Remote Sens. Environ. 185, 57–70. https://doi.org/10.1016/j.rse.2015.12.024.

Saintilan, N., Wilson, N.C., Rogers, K., Rajkaran, A., Krauss, K.W., 2014. Mangrove expansion and salt marsh decline at mangrove poleward limits. Glob. Chang. Biol. 20, 147–157. https://doi.org/10.1111/gcb.12341.

Sitch, S., Friedlingstein, P., Gruber, N., Jones, S.D., Murray-Tortarolo, G., Ahlström, A., Doney, S.C., Graven, H., Heinze, C., Huntingford, C., Levis, S., Levy, P.E., Lomas, M., Poulter, B., Viovy, N., Zaehle, S., Zeng, N., Arneth, A., Bonan, G., Bopp, L., Canadell, J.G., Chevallier, F., Ciais, P., Ellis, R., Gloor, M., Peylin, P., Piao, S.L., Le Quéré, C., Smith, B., Zhu, Z., Myneni, R., 2015. Recent trends and drivers of regional sources and sinks of carbon dioxide. Biogeosciences. https://doi.org/10.5194/bg-12-653-2015.

Taylor, S.J., Letham, B., 2018. Forecasting at Scale. Am. Stat. 72, 37–45. https://doi.org/10.1080/00031305.2017.1380080.

Taylor, S.J., Letham, B., Taylor, S.J., Letham, B., 2018. Forecasting at Scale. Am. Stat. 72, 37–45. https://doi.org/10.1080/00031305.2017.1380080.

Tomlinson, P.B., 1986. The botany of mangroves.

van Bussel, L.G.J., Ewert, F., Leffelaar, P.A., 2011. Effects of data aggregation on simulations of crop phenology. Agric. Ecosyst. Environ. 142, 75–84. https://doi.org/10.1016/j.agee.2010.03.019.

Verbesselt, J., Hyndman, R., Newnham, G., Culvenor, D., 2010. Detecting trend and seasonal changes in satellite image time series. Remote Sens. Environ. 114, 106–115. https://doi.org/10.1016/j.rse.2009.08.014.

Vogelmann, J.E., Gallant, A.L., Shi, H., Zhu, Z., 2016. Perspectives on monitoring gradual change across the continuity of Landsat sensors using time-series data. Remote Sens. Environ. https://doi.org/10.1016/j.rse.2016.02.060.

White, M.A., De Beurs, K.M., Didan, K., Inouye, D.W., Richardson, A.D., Jensen, O.P., O'keefe, J., Zhang, G., Nemani, R.R., Van Leeuwen, W.J.D., Brown, J.F., De Wit, A., Schaepman, M., Lin, X., Dettinger, M., Bailey, A.S., Kimball, J., Schwartz, M.D., Baldocchi, D.D., Lee, J.T., Lauenroth, W.K., 2009. Intercomparison, interpretation, and assessment of spring phenology in North America estimated from remote sensing for 1982–2006. Glob. Chang. Biol. 15, 2335–2359. https://doi.org/10.1111/j.1365-2486.2009.01910.x.

White, M.A., Nemani, R.R., Thornton, P.E., Running, S.W., 2002. Satellite evidence of phenological differences between urbanized and rural areas of the eastern United States deciduous broadleaf forest. Ecosystems 5, 260–273. https://doi.org/10.1007/s10021-001-0070-8.

Wood, S., 2017. Generalized Additive Models An Introduction with R, Second Edi. ed. Chapman and Hall/CRC, NEW YORK. https://doi.org/10.1201/978131537027.

Wu, H., Li, Z.-L.L., 2009. Scale issues in remote sensing: A review on analysis, processing and modeling. Sensors 9, 1768–1793. https://doi.org/10.3390/s90301768.

Wulder, M.A., White, J.C., Loveland, T.R., Woodcock, C.E., Belward, A.S., Cohen, W.B., Fosnight, E.A., Shaw, J., Masek, J.G., Roy, D.P., 2016. The global Landsat archive: Status, consolidation, and direction. Remote Sens. Environ. 185, 271–283. https://doi.org/10.1016/j.rse.2015.11.032.

Yan, J., Wang, L., Song, W., Chen, Y., Chen, X., Deng, Z., 2019. A time-series classification approach based on change detection for rapid land cover mapping. ISPRS J. Photogramm. Remote Sens. 158, 249–262. https://doi.org/10.1016/j.isprsjprs.2019.10.003.

Younes, N., Joyce, K.E., Northfield, T.D., Maier, S.W., 2019. The effects of water depth on estimating Fractional Vegetation Cover in mangrove forests. Int. J. Appl. Earth Obs. Geoinf. 83, 101924 https://doi.org/10.1016/j.jag.2019.101924.

Zhang, H.K., Roy, D.P., 2016. Landsat 5 Thematic Mapper reflectance and NDVI 27-year time series inconsistencies due to satellite orbit change. Remote Sens. Environ. 186, 217–233. https://doi.org/10.1016/j.rse.2016.08.022.

Zhou, Q., Rover, J., Brown, J., Worstell, B., Howard, D., Wu, Z., Gallant, L.A., Rundquist, B., Burke, M., 2019. Monitoring Landscape Dynamics in Central U.S. Grasslands with Harmonized Landsat-8 and Sentinel-2 Time Series Data. Remote Sens. https://doi.org/10.3390/rs11030328.

Zuur, A.F., Saveliev, A.A., Ieno, E.N., 2014. A beginner's guide to generalised additive mixed models with R.